Structural and Copy Number Variants

Mikhail Dozmorov

2021-05-03

SVs - structural variants

- Structural variation (SV) is defined as differences in the copy number, orientation or location of relatively large genomic segments (typically $>$100 bp).
- Two humans differ by 5,000–10,000 inherited SVs
- Both inherited and *de novo* SVs contribute to a variety of normal and disease phenotypes

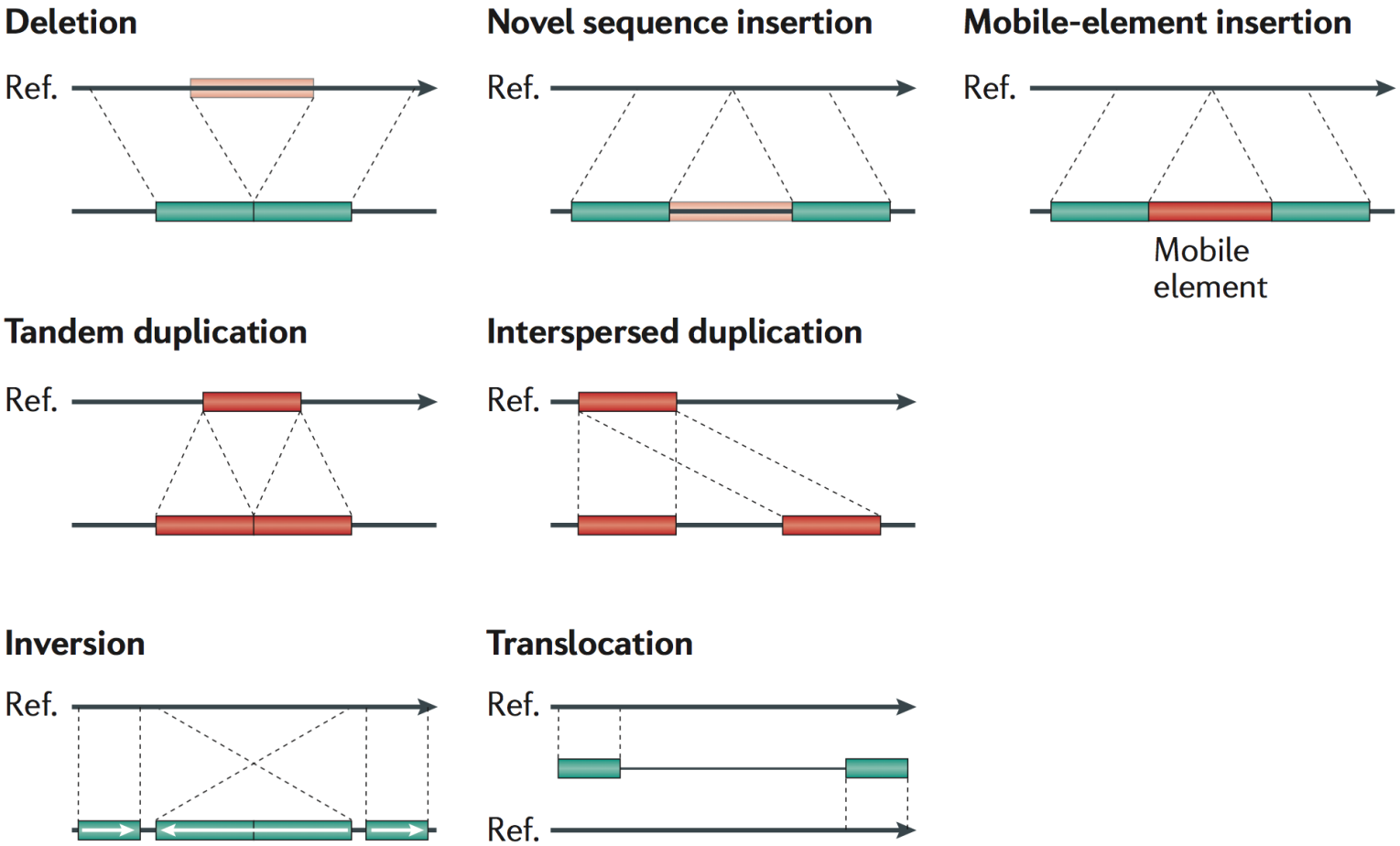
SV

- **Structural variant (SV)** - Genomic rearrangements that affect >1kb of sequence, including deletions, novel insertions, inversions, mobile-element transpositions, duplications and translocations.
- **Copy number variant (CNV)** - Unbalanced structural variants; variants that change the number of base pairs in the genome.
- **Mobile elements** - DNA sequences that move location within the genome. Active mobile elements (transposons) in the human genome include Alu, L1 and SVA sequences.

Large CNVs are individually very rare in the general population, yet 8% of individuals have a CNV of >500 kb in their genomes

SV types

SVs vary widely in size and there are numerous classes of structural variation: deletions, translocations, inversions, mobile elements, tandem duplications and novel insertions



Why is structural variation relevant / important?

- They are common and affect more base pairs than all single-nucleotide differences.
- Genetic basis of traits
  - Cancer - Genome instability and aneuploidy
  - Neuropsychiatric disease (e.g., autism, schizophrenia)
  - Developmental disorders - Spontaneous SVs have been implicated in so-called “genomic” and developmental disorders
  - Age-dependent disease - Somatic genome instability

SV and human disease phenotypes

Table 2 Examples of copy number variations (CNVs) and conveyed genomic disorders[a]

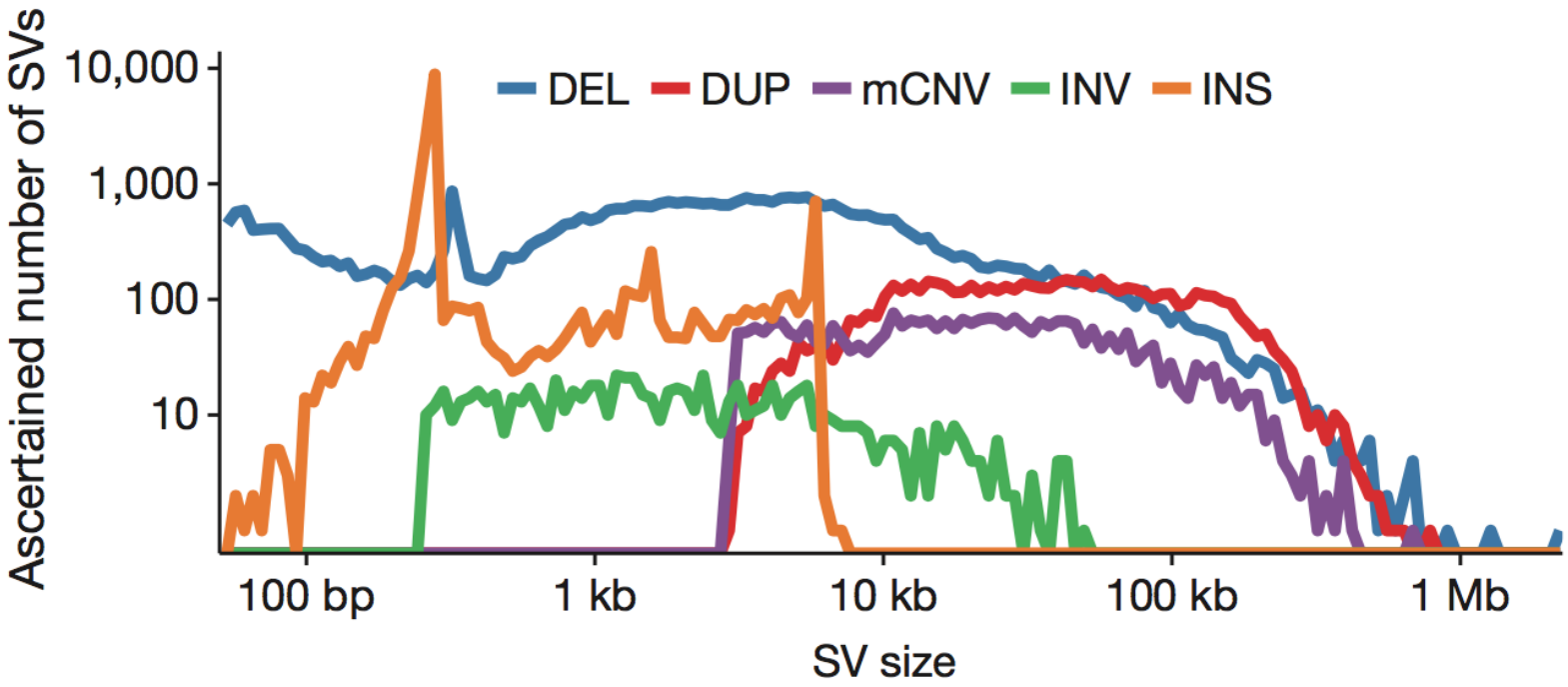
Phenotype	OMIM	Locus	CNV	References[a]
Hunter syndrome	309900	*IDS*	del/inv	S8, S70, S72
Ichthyosis	308100	*STS*	del	S56
Mental retardation	300706	*HUWE1*	dup	S21
Pelizaeus-Merzbacher disease	312080	*PLP1*	del/dup/tri	S14, S28, S37, S38, S71
Progressive neurological symptoms (MR+SZ)	300260	*MECP2*	dup	S3, S15, S65
Red-green color blindness	303800	opsin genes	del	S46
Complex traits				
Alzheimer disease	104300	*APP*	dup	S52
Autism	612200	3q24	inherited homozygous del	S45
	611913	16p11.2	del/dup	S34, S42, S54, S68
Crohn disease	266600	*HBD-2*	copy number loss	S20
	612278	*IRGM*	del	S44
HIV susceptibility	609423	*CCL3L1*	copy number loss	S23, S33
Mental retardation	612001	15q13.3	del	S58
	610443	17q21.31	del	S32, S57, S59
	300534	Xp11.22	dup	S21
Pancreatitis	167800	*PRSS1*	tri	S36
Parkinson disease	168600	*SNCA*	dup/tri	S12, S19, S22, S27, S61

(*Continued*)

http://www.annualreviews.org/doi/full/10.1146/annurev.genom.9.081307.164217

Why is structural variation relevant / important?

Size distribution of SVs in 1000 genomes project



Sudmant, Peter H., Tobias Rausch, Eugene J. Gardner, Robert E. Handsaker, Alexej Abyzov, John Huddleston, Yan Zhang, et al. "An Integrated Map of Structural Variation in 2,504 Human Genomes." Nature 526, no. 7571 (September 30, 2015): 75–81. https://doi.org/10.1038/nature15394.

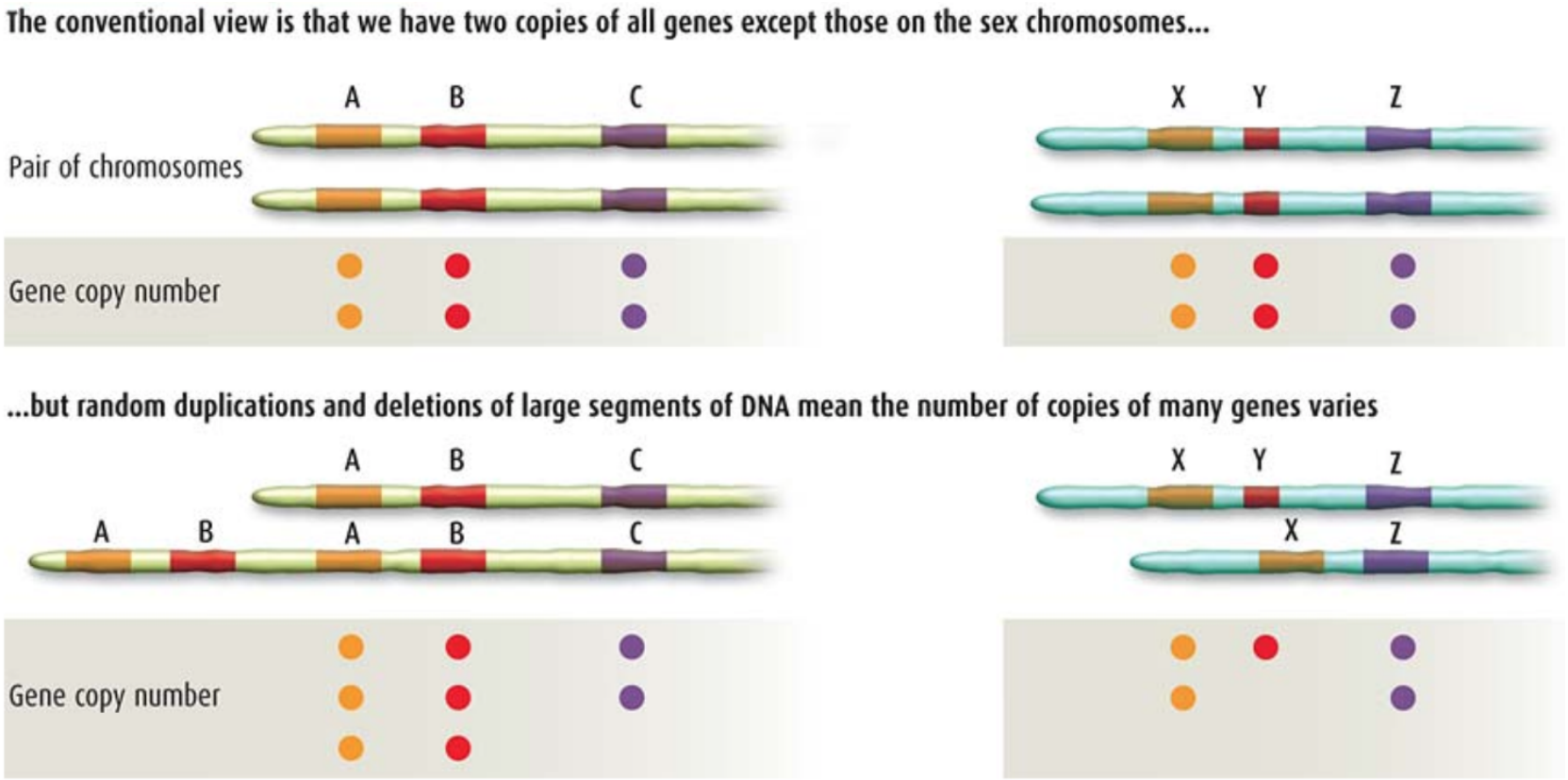
Copy Number Variants (CNVs)

- Copy number variants (deletions/duplications $>$ 50 bp) account for more inter-individual variation than do single-nucleotide variants
- In an average haploid human sequence,
  - ~9 Mb are affected by structural variants,
  - ~3.6 Mb are affected by single nucleotide variants,
  - on average, humans are heterozygous for ~150 CNVs (Sudmant et al., 2015, Nature)

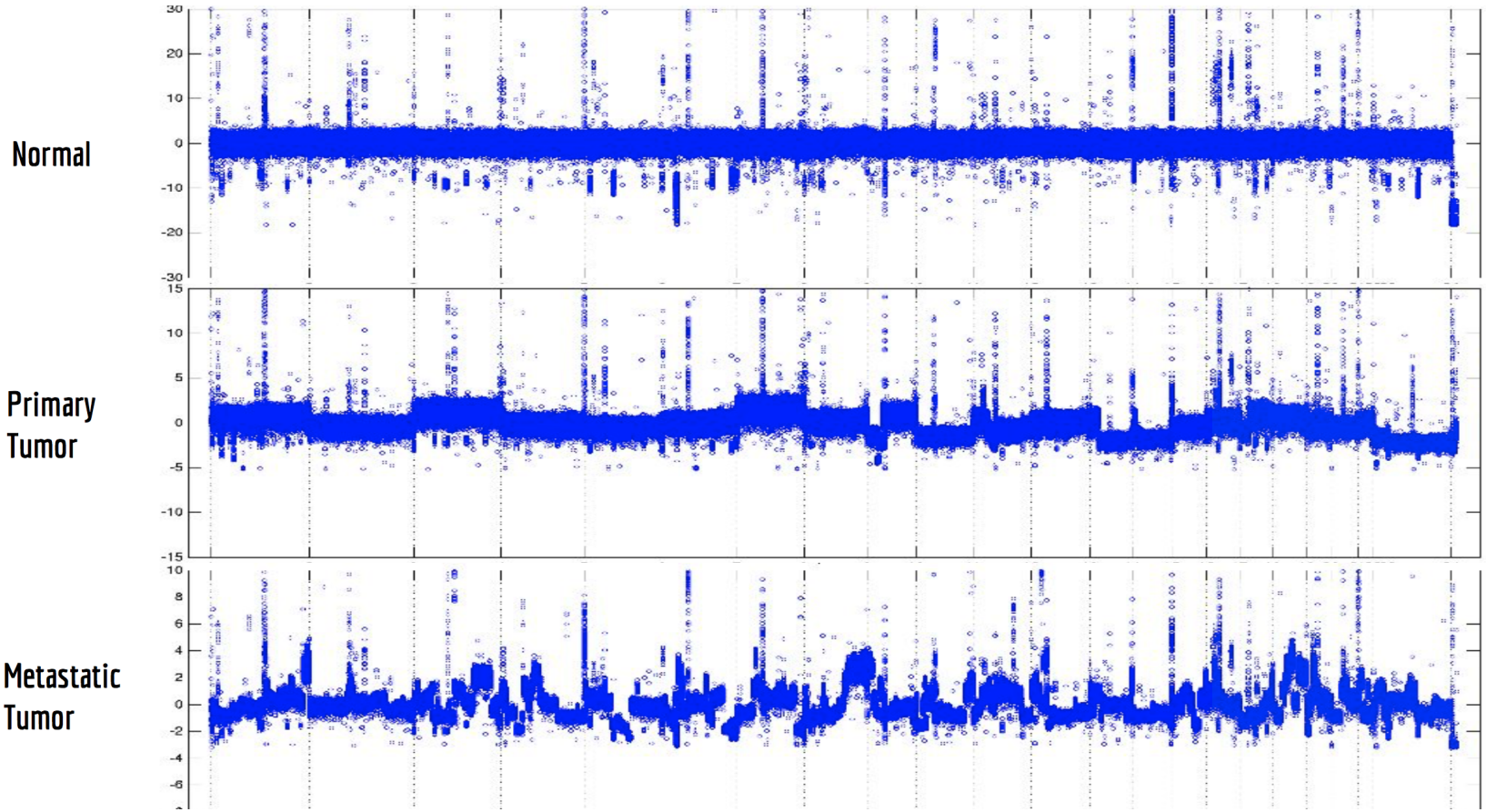
Sudmant, Peter H., Tobias Rausch, Eugene J. Gardner, Robert E. Handsaker, Alexej Abyzov, John Huddleston, Yan Zhang, et al. "An Integrated Map of Structural Variation in 2,504 Human Genomes." Nature 526, no. 7571 (September 30, 2015): 75–81. https://doi.org/10.1038/nature15394.

Iafrate, A John, Lars Feuk, Miguel N Rivera, Marc L Listewnik, Patricia K Donahoe, Ying Qi, Stephen W Scherer, and Charles Lee. "Detection of Large-Scale Variation in the Human Genome." Nature Genetics 36, no. 9 (September 2004): 949–51. doi:10.1038/ng1416.

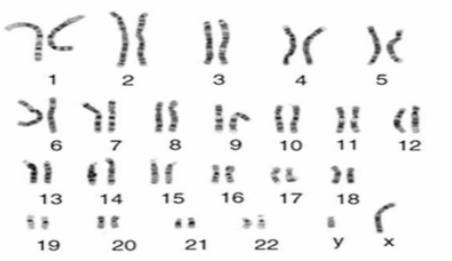
Copy Number Variants (CNVs)



CNVs in tumors

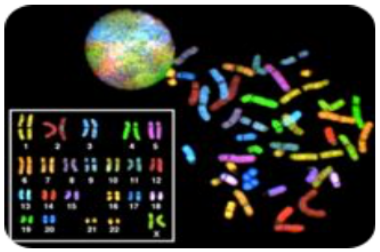


Technologies assessing genome stability



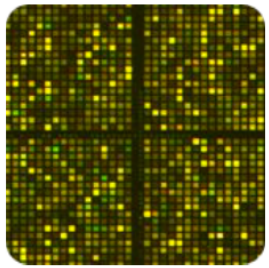
1940s - 1980s

Cytogenetics / Karyotyping



1990s

CGH / FISH /
SKY / COBRA



2000s

Genomic microarrays
BAC-aCGH / oligo-aCGH

Today

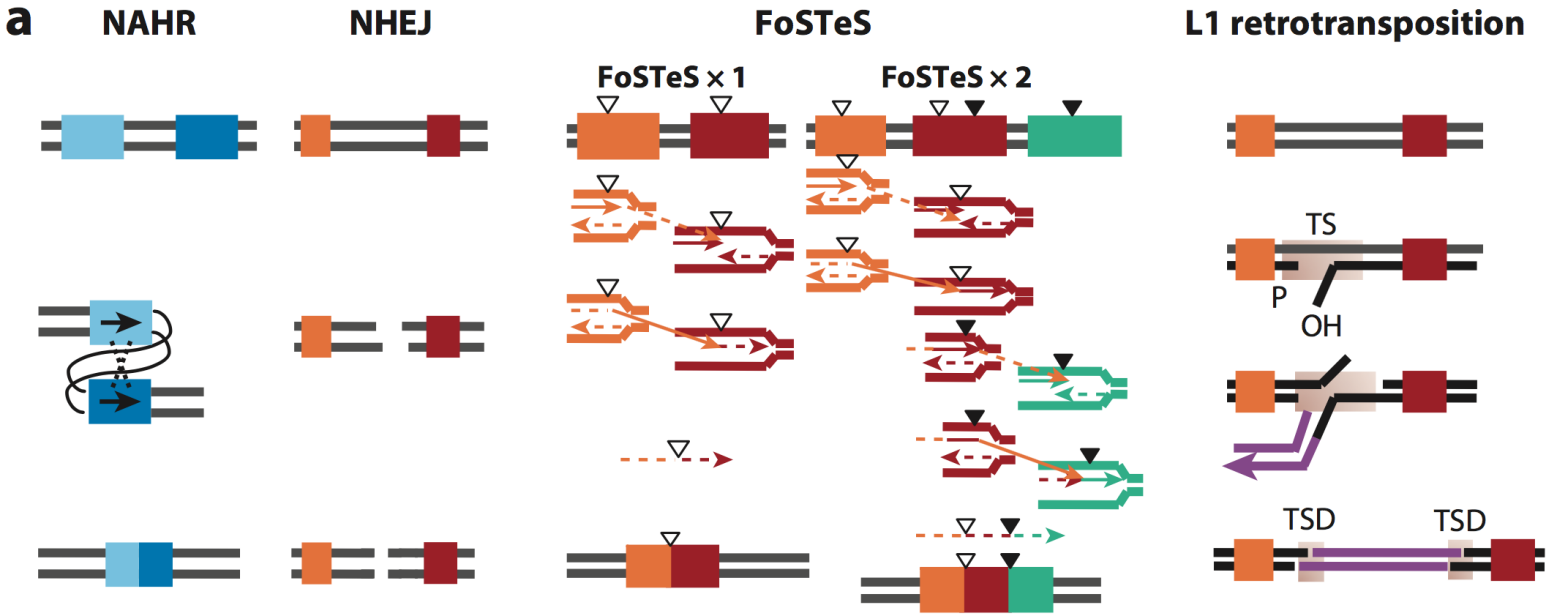
High throughput
DNA sequencing

How CNVs arise?

Four major mechanisms:

- **NAHR** - Non-Allelic Homologous Recombination between repeat sequences
- **NHEJ** - Non-Homologous End-Joining, recombination repair of double strand break
- **FoSTeS** - Fork Stalling and Template Switching. Multiple FoSTeS events ($\times 2$ or more) result in complex rearrangements, single FoSTeS event ($\times 1$) cause simple rearrangements
- **L1-mediated retrotransposition**

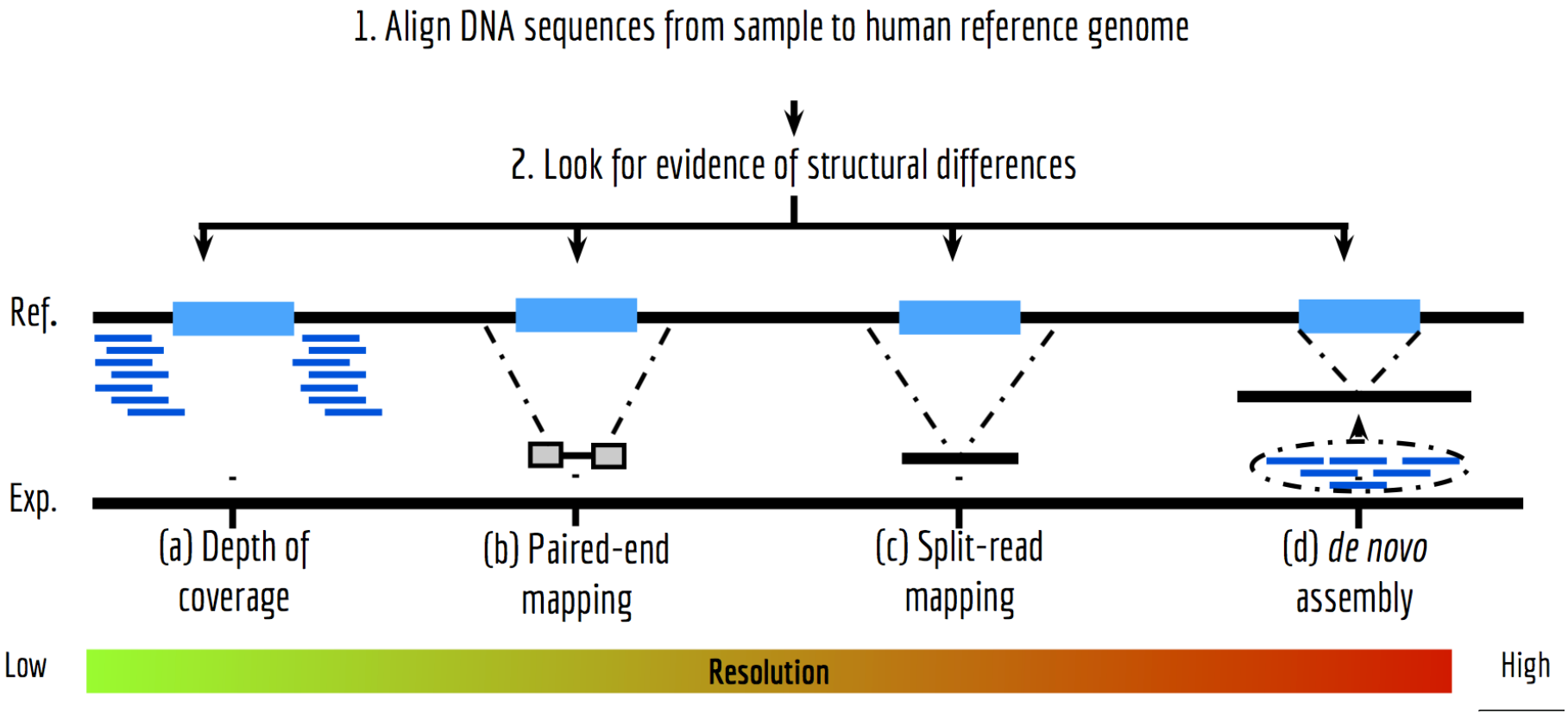
Mechanisms



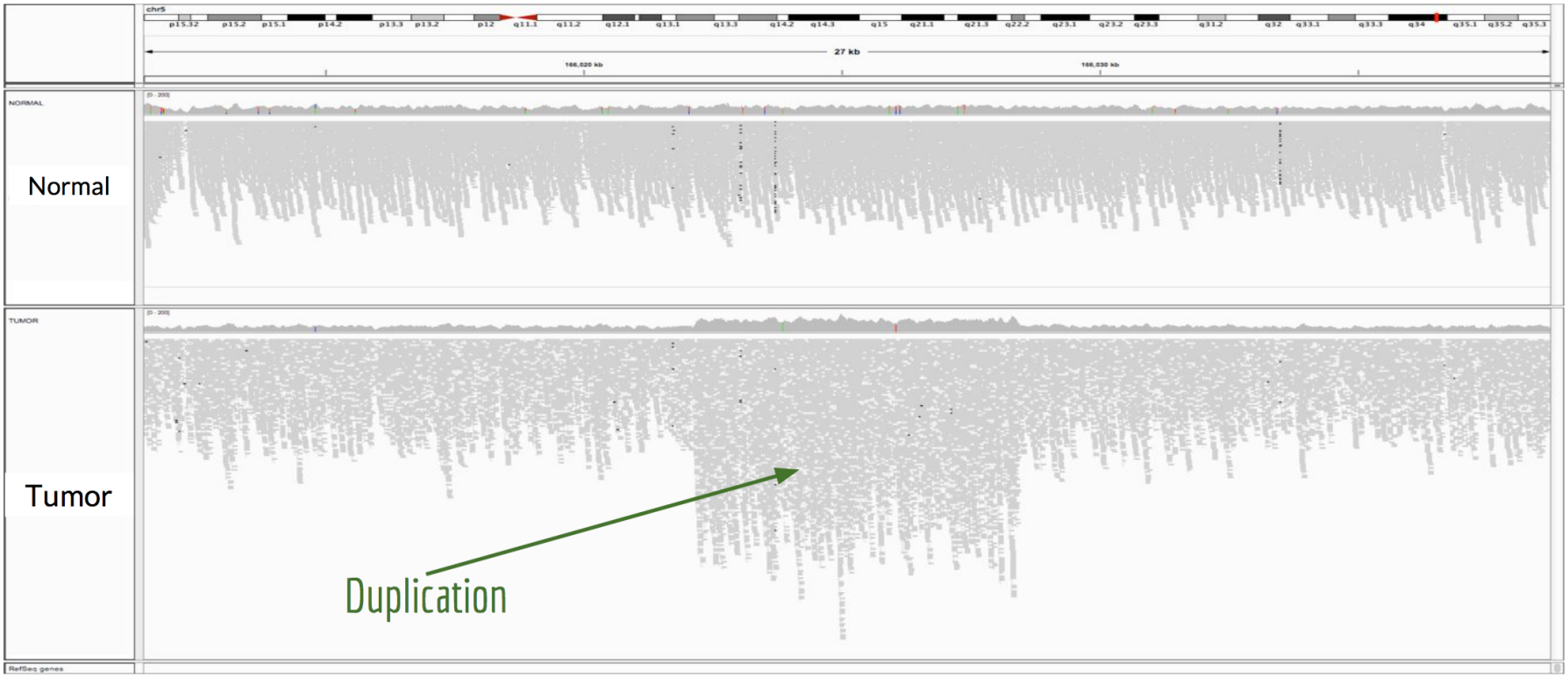
b	NAHR	NHEJ	FoSTeS	Retrotransposition
Structural variation type	dup, del, inv	dup, del	dup, del, inv, complex	ins
Homology flanking breakpoint (before rearrangement)?	Yes (LCR/SD, *Alu*, L1, or pseudogene)	No	No	No
Breakpoint	Inside homology	Addition or deletion of basepairs, or microhomology	Microhomology	No specification
Sequence undergoing SV	Any	Any	Any	Transcribed sequences

Zhang, Feng, Wenli Gu, Matthew E. Hurles, and James R. Lupski. "Copy Number Variation in Human Health, Disease, and Evolution." Annual Review of Genomics and Human Genetics 10 (2009): 451–81.

How do we identify structural variants via DNA sequencing?



Copy number affect the depth of sequence coverage



Challenges:

- need high coverage for high resolution
- deletions easier than duplications
- prone to artifacts owing to repeats, GC content, etc.

Detecting CNV by counting alignments in genome "windows"

Strengths:

1) Fast and simple.
2) Easy to identify gene amplifications.
3) Relatively straightforward interpretation: is gene X amplified or deleted?

Weaknesses:

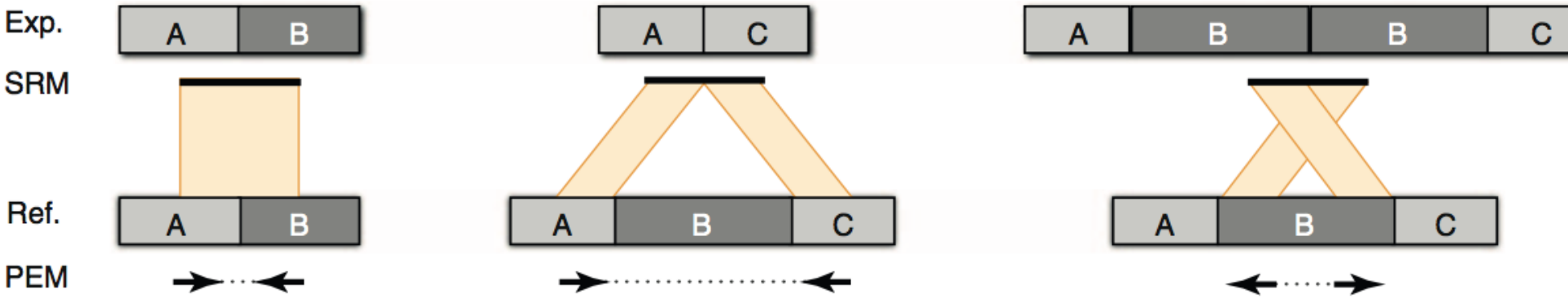
1) Limited resolution (2-5kb) = imprecise boundaries
2) Cannot detect balanced events or reveal variant architecture.

Best practices

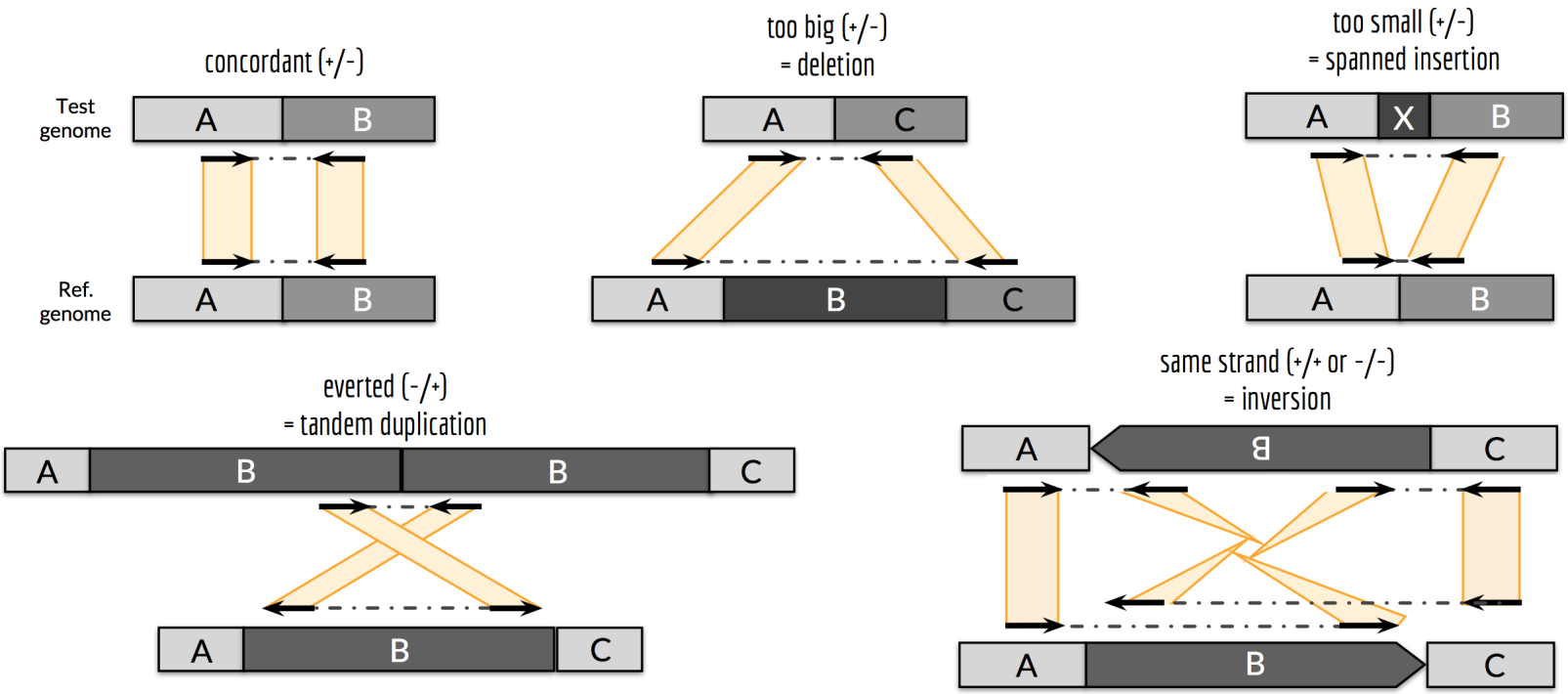
- Use variably-sized windows, masked for repeats - repeatMasker, simple sequence repeats, mappability
- Window size should yield >100 reads (median)

Paired-end mapping

- Sequencing libraries are created with fragments of known length (generally 200–500 bp for paired-end libraries and 1–10 kb for mate-pair libraries).
- Paired-end sequences that are 'concordant' with the reference genome align with the expected distance and orientation
- Read pairs spanning an SV breakpoint will produce 'discordant' alignments with an unexpected alignment distance and/or orientation.



Discordant mapping "signatures" for various SV types



Quinlan, Aaron R., and Ira M. Hall. "Characterizing Complex Structural Variation in Germline and Somatic Genomes." Trends in Genetics: TIG 28, no. 1 (January 2012): 43–53. https://doi.org/10.1016/j.tig.2011.10.002.

Looking for "discordant" paired-end fragments

Challenges:

- Difficult to achieve single-nucleotide resolution for the SV breakpoint
- Chimeric molecules, PCR duplicates

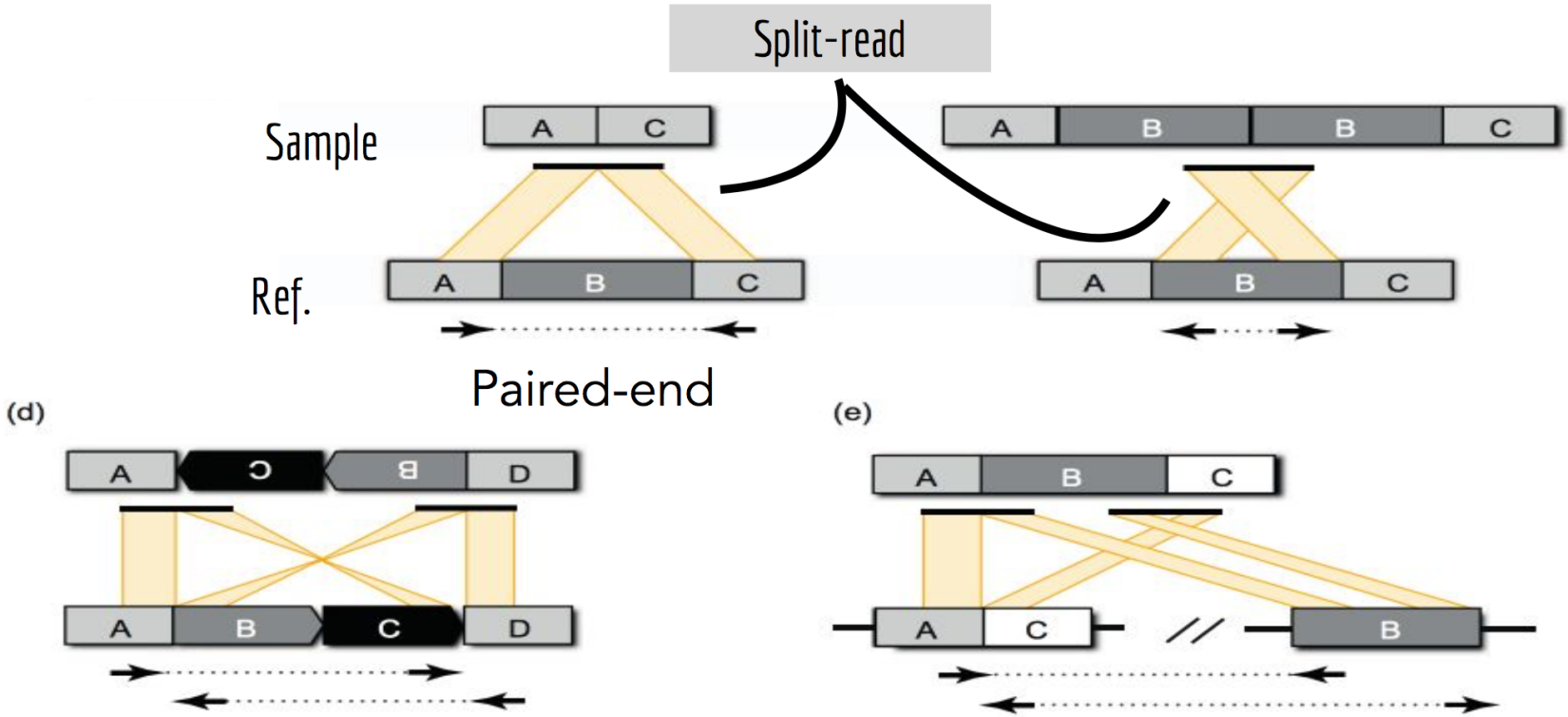
Advantages:

- Much higher resolution
- Can find any type of SV - not limited to deletions and duplications like depth of coverage

Split-read mapping

- SRM identifies sequences that actually contain a breakpoint
- The alignments for such sequences are 'split' because DNA segments flanking the breakpoint align to disjoint locations in the reference genome.
- SRM inherently maps breakpoints to single base resolution
- SRM requires reads longer than approximately 200 bp. Long-read (>500 bp) SRM is a particularly powerful approach for studying complex SV because multiple breakpoints can potentially be captured by a single read

Split-read mapping "signatures" for various SV types



Tools: Read-depth methods

Read-depth approaches assume a random (typically Poisson or modified Poisson) distribution in mapping depth and investigate the divergence from this distribution to discover duplications and deletions in the sequenced sample.

The basic idea is that duplicated regions will show significantly higher read depth and deletions will show reduced read depth when compared to diploid regions

- **CNVnator** - a tool for CNV discovery and genotyping from depth of read mapping., http://sv.gersteinlab.org/
- **AGE** - a tools that implements an algorithm for optimal alignment of sequences with SVs, http://sv.gersteinlab.org/

Tools: Split-read approaches

Split-read methods are capable of detecting deletions and small insertions down to single-base-pair resolution and were first applied to longer Sanger sequencing reads.

The aim is to define the breakpoint of a structural variant on the basis of a 'split' sequence-read signature (that is, the alignment to the genome is broken; a continuous stretch of gaps in the read indicates a deletion or in the reference indi- cates an insertion).

- **Pindel** - can detect breakpoints of large deletions, medium sized insertions, inversions, tandem duplications and other structural variants at single-based resolution from next-gen sequence data. It uses a pattern growth approach to identify the breakpoints of these variants from paired-end short reads.
  http://gmt.genome.wustl.edu/packages/pindel/

Tools: Paired-end mapping

Read-pair methods assess the span and orientation of paired-end reads and cluster 'discordant' pairs in which the mapping span and/or orientation of the read pairs are inconsistent with the reference genome

Read pairs that map too far apart define deletions, those found too close together are indicative of insertions, and orientation inconsistencies can delineate inversions and a specific class of tandem duplications

- **PEMer** - http://sv.gersteinlab.org/pemer/
- **VariationHunter** - http://variationhunter.sourceforge.net/Home
- **BreakDancer** - http://breakdancer.sourceforge.net/
- **MoDIL** - http://compbio.cs.toronto.edu/modil/
- **HydraMulti** - an SV discovery tool that incorporates hundreds of samples, https://github.com/arq5x/Hydra
- **Spanner** - Spanner is a c++ program for the detection of Structural Variation events from whole genome sequenced read pair data. https://github.com/chipstewart/Spanner

Tools: Sequence assembly

In theory, all forms of structural variation could be accurately typed for copy, content and structure if the underlying sequence reads were long and accurate enough to allow de novo assembly. In practice, sequence-assembly approaches are still in their infancy and typically use a combination of de novo and local assembly algorithms to generate sequence contigs that are then compared to a reference genome

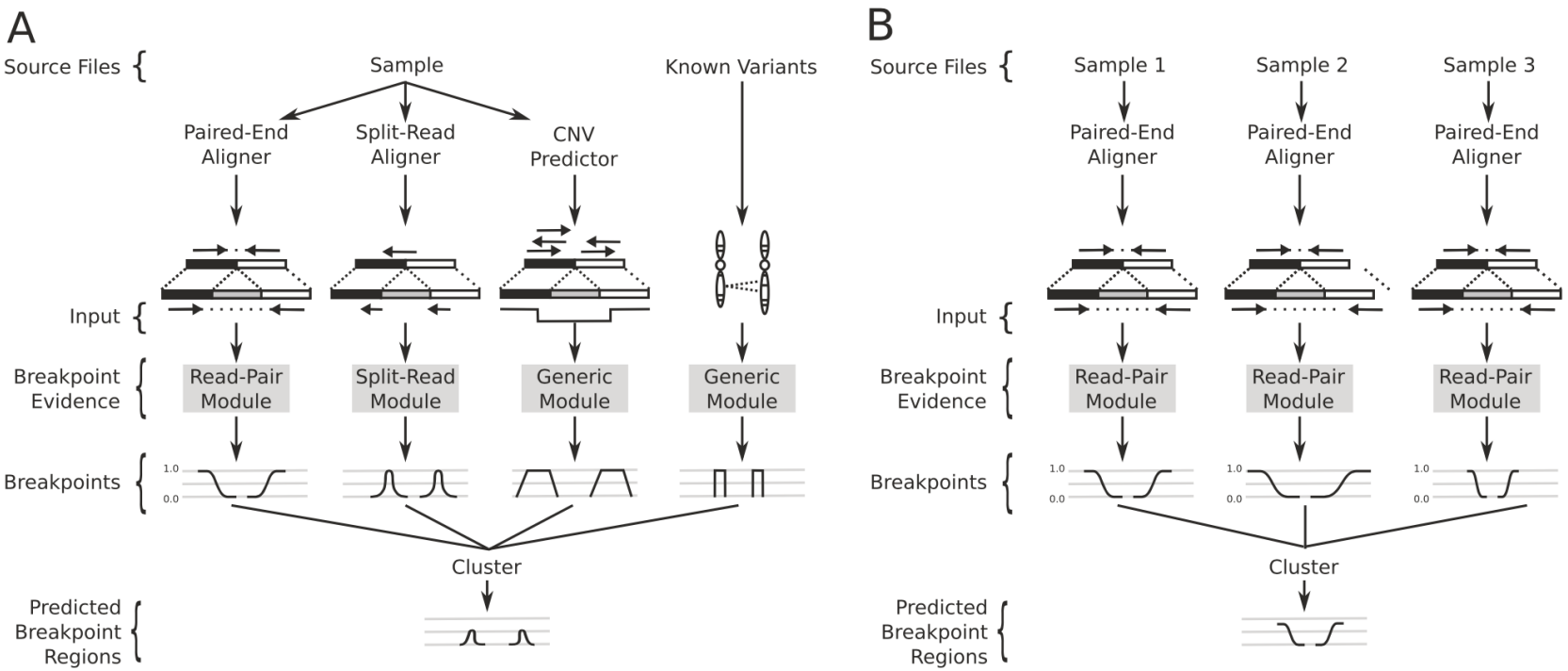
- **SOAPdenovo** - http://soap.genomics.org.cn/soapdenovo.html
- **ALLPATH-LG** - http://software.broadinstitute.org/allpaths-lg/blog/
- **Cortex** - http://cortexassembler.sourceforge.net/
- **NovelSeq** - http://compbio.cs.sfu.ca/software-novelseq
- **TIGRA** - http://bioinformatics.mdanderson.org/main/TIGRA

Tools: Other approaches

- **DELLY2** - Structural variant discovery by integrated paired-end and split-read analysis. https://github.com/dellytools/delly
- **Genome STRiP** (Genome STRucture In Populations) is a suite of tools for discovering and genotyping structural variations using sequencing data. The methods are designed to detect shared variation using data from multiple individuals.
  http://software.broadinstitute.org/software/genomestrip/

Tools: Other approaches

- **LUMPY-SV** - a general probabilistic framework for structural variant discovery. Integrates multiple signals - read-pair, split-read, read-depth and prior knowledge. Operates on multiple samples.

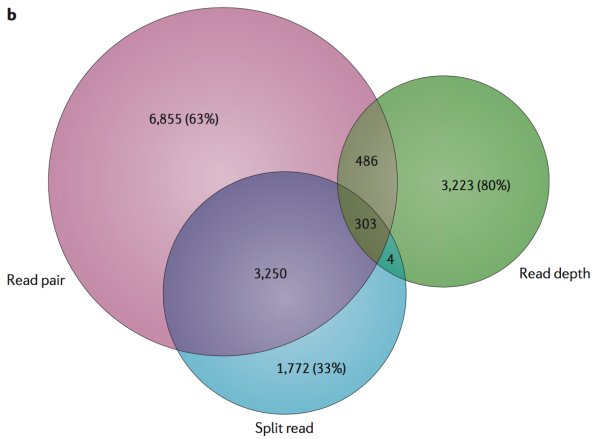


https://github.com/arq5x/lumpy-sv/

SV detection methods summary

Limitations

- On the basis of typical NGS fragment sizes, more than 90% of the discovered events are less than 1 kb and most of these are deletions rather than insertions
- Over 1.5% of the human genome cannot be covered uniquely even with read lengths of 1 kb
- Low reproducibility

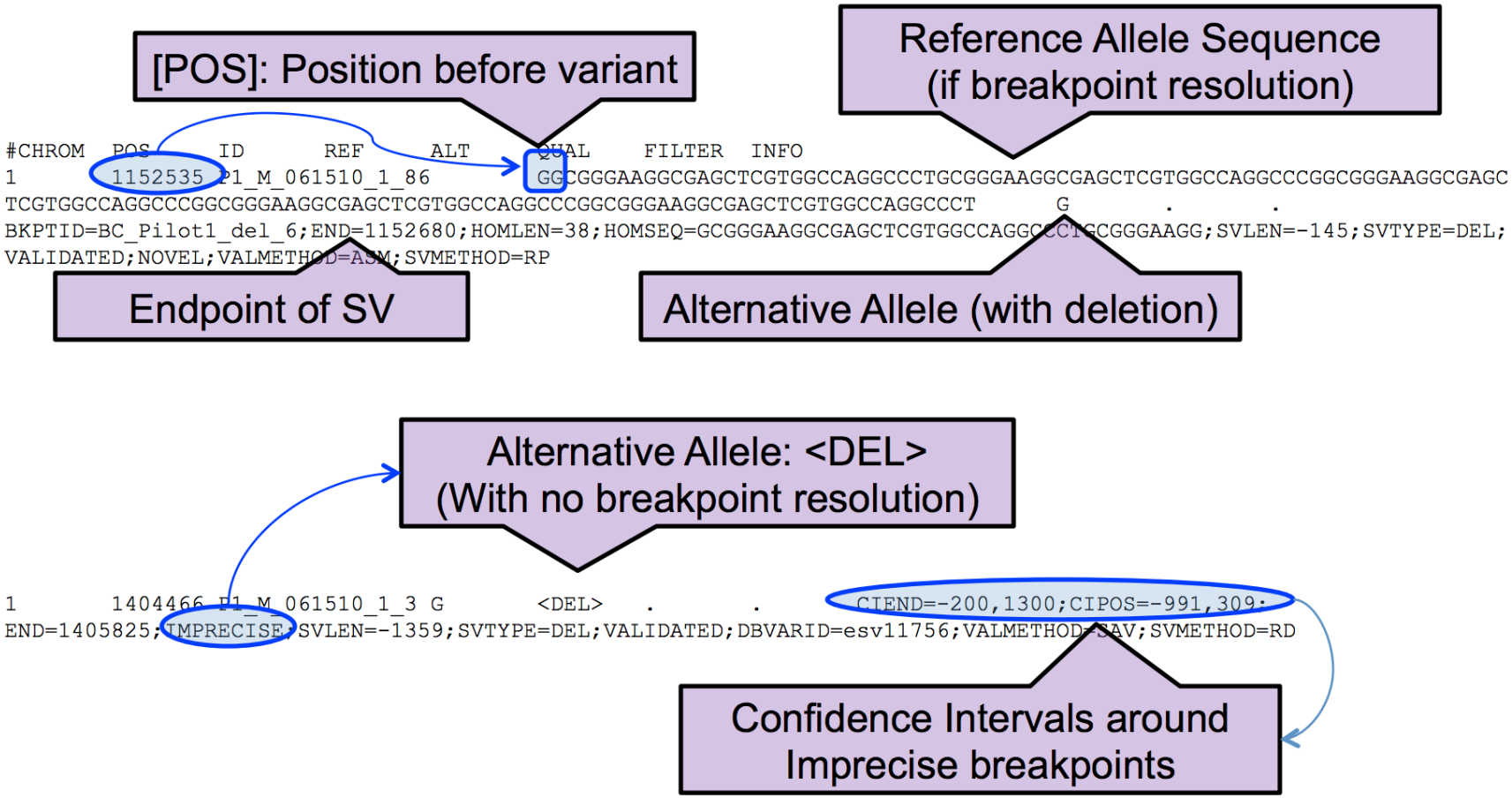


The most serious challenges that remain are the absence of a 'gold standard' for assessment of disparate discovery and genotyping methods, and the remaining biases in global discovery

SV discovery set in VCF format

- VCF Format
  - `#CHROM POS ID REF ALT QUAL FILTER INFO`
  - [POS] is the position before the variant
  - [ID] links the variant to the original SV discovery method and callset (SV master validation tables)
  - [REF] and [ALT] show exact sequence if breakpoints are known, otherwise a variant-specific tag is used: (, , , ,)
  - [INFO] contains various information including [END] as the SV end coordinate
- Processed with `vcftools`: http://vcftools.sourceforge.net/

Example VCF Records for SVs



Processing VCF genotypes with vcftools

- `--012` converts vcf file into large matrix with samples as columns and genotypes as 0,1,2 representing the number of non-reference alleles
- `--IMPUTE` converts vcf file into IMPUTE reference-panel format
- `--BEAGLE-GL` converts vcf into input file for the BEAGLE program
- `--plink` converts vcf into PLINK PED format

Full list of commands can be found here:
http://vcftools.sourceforge.net/options.html

Problems in SV calling

- Often many false positives (~30%)
- Short reads + heuristic alignment + rep. genome = systematic alignment artifacts (false calls)
- Chimeras and duplicate molecules
- Ref. genome errors (e.g., gaps, mis-assemblies)
- ALL SV mapping studies use strict filters for above

Solution - long-read sequencing

LETTER

doi:10.1038/nature13907

Resolving the complexity of the human genome using single-molecule sequencing

Mark J. P. Chaisson[1], John Huddleston[1,2], Megan Y. Dennis[1], Peter H. Sudmant[1], Maika Malig[1], Fereydoun Hormozdiari[1], Francesca Antonacci[3], Urvashi Surti[4], Richard Sandstrom[1], Matthew Boitano[5], Jane M. Landolin[5], John A. Stamatoyannopoulos[1], Michael W. Hunkapiller[5], Jonas Korlach[5] & Evan E. Eichler[1,2]

The human genome is arguably the most complete mammalian reference assembly[1–3], yet more than 160 euchromatic gaps remain[4–6] and aspects of its structural variation remain poorly understood ten years after its completion[7–9]. To identify missing sequence and genetic variation, here we sequence and analyse a haploid human genome (CHM1) using single-molecule, real-time DNA sequencing[10]. We close or extend 55% of the remaining interstitial gaps in the human GRCh37 reference genome—78% of which carried long runs of degenerate short tandem repeats, often several kilobases in length, embedded within (G+C)-rich genomic regions. We resolve the complete sequence of 26,079 euchromatic structural variants at the base-pair level, including inversions, complex insertions and long tracts of tandem repeats. Most have not been previously reported, with the greatest increases in sensitivity occurring for events less than 5 kilobases in size. Compared to the human reference, we find a significant insertional bias (3:1) in regions corresponding to complex insertions and long short tandem repeats. Our results suggest a greater complexity of the human genome in the form of variation of longer and more complex repetitive DNA that can now be largely resolved with the application of this longer-read sequencing technology.

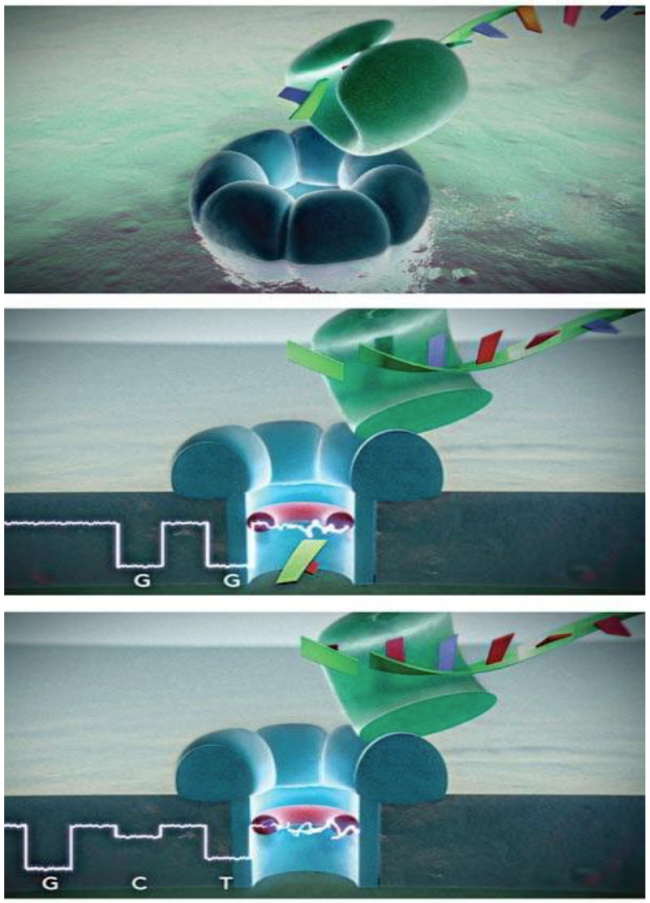
Data generated by single-molecule, real-time (SMRT) sequencing technology differ drastically from most sequencing platforms because native DNA is sequenced without cloning or amplification, and read

for recruiting additional sequence reads for assembly (Supplementary Information). Using this approach, we closed 50 gaps and extended into 40 others (60 boundaries), adding 398 kb and 721 kb of novel sequence to the genome, respectively (Supplementary Table 4). The closed gaps in the human genome were enriched for simple repeats, long tandem repeats, and high (G+C) content (Fig. 1) but also included novel exons (Supplementary Table 20) and putative regulatory sequences based on DNase I hypersensitivity and chromatin immunoprecipitation followed by high-throughput DNA sequencing (ChIP-seq) analysis (Supplementary Information). We identified a significant 15-fold enrichment of short tandem repeats (STRs) when compared to a random sample ($P < 0.00001$) (Fig. 1a). A total of 78% (39 out of 50) of the closed gap sequences were composed of 10% or more of STRs. The STRs were frequently embedded in longer, more complex, tandem arrays of degenerate repeats reaching up to 8,000 bp in length (Extended Data Fig. 1a–c), some of which bore resemblance to sequences known to be toxic to *Escherichia coli*[16]. Because most human reference sequences[17,18] have been derived from clones propagated in *E. coli*, it is perhaps not surprising that the application of a long-read sequence technology to uncloned DNA would resolve such gaps. Moreover, the length and complex degeneracy of these STRs embedded within (G+C)-rich DNA probably thwarted efforts to follow up most of these by PCR amplification and sequencing.

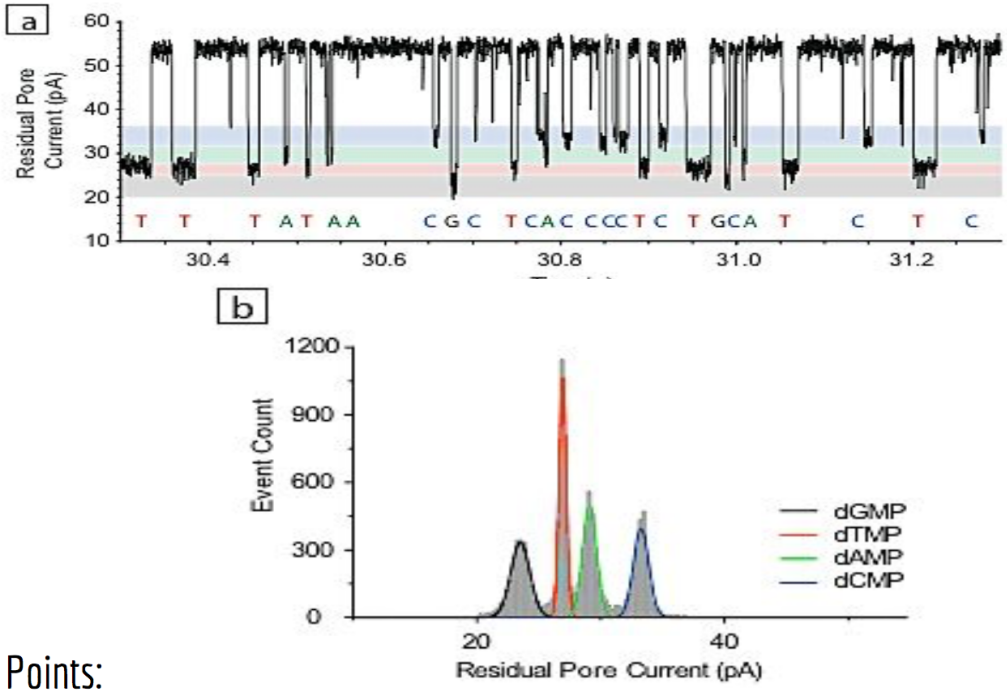
Next, we developed a computational pipeline (Extended Data Fig. 2)

https://www.nature.com/nature/journal/v517/n7536/full/nature13907.html

Oxford Nanopore Sequencing



Clarke et al., 2009: Nature Nanotechnology



Key Points:

- Protein nanopore array embedded in an artificial lipid
- 1 DNA molecule, 1 translocating enzyme
- salt + electrodes on either side of pore
- Bases detected by change in current
- intrinsic detection of methylated cytosine

Long read analysis

- `poretools` - a toolkit for working with Oxford nanopore data

Table 1. Summary of currently supported operations in poretools

Command	Description
`combine`	Combine a set of FAST5 files in a TAR archive.
`events`	Extract each nanopore event for each read.
`fasta`	Extract FASTA sequences from a set of FAST5 files.
`fastq`	Extract FASTQ sequences from a set of FAST5 files.
`hist`	Plot read size histogram for a set of FAST5 files.
`nucdist`	Measure the nucleotide composition.
`qualdist`	Measure the quality score composition.
`readstats`	Extract signal information for each read over time.
`squiggle`	Plot the observed signals for FAST5 reads.
`stats`	Get read size stats for a set of FAST5 files.
`tabular`	Extract sequence information in TAB delimited format
`times`	Return the start times from a set of FAST5 files.
`winner`	Extract the longest read from a set of FAST5 files.
`yield_plot`	Plot the sequencing yield over time.

SpeedSeq genome analysis pipeline

- Integrates **FreeBayes**, **LUMPY** for breakpoint detection, **SVTyper**

SVTyper

- SVTyper is a maximum-likelihood Bayesian classification algorithm that infers an underlying genotype at each SV
- $S(g)$ is the prior probability of observing a variant read in a single trial given a genotype g at any locus
- Assuming a random sampling of reads, the number of observed alternate (A) and reference (R) reads will follow a binomial distribution $B(A + R, S(g'))$, where $g' \in G$ is the true underlying genotype

SVTyper

$$S(g) = \begin{cases} 0.1 & \text{if } g = \textit{homozygous refererence} \\ 0.4 & \text{if } g = \textit{heterozygous} \\ 0.8 & \text{if } g = \textit{homozygous alternate} \end{cases}$$

$$P(A, R \mid g) = \binom{A+R}{A} \cdot S(g)^A \cdot (1 - S(g))^R$$

SVTyper

$$P(g \mid A,R) = \frac{P(A,R \mid g) \cdot P(g)}{P(A,R)} = \frac{P(A,R \mid g) \cdot P(g)}{\sum_{g \in G} P(A,R \mid g) \cdot P(g)}$$

$$\hat{g} = \arg\max_{g \in G} P(g \mid A,R)$$

Chiang, Colby, Ryan M Layer, Gregory G Faust, Michael R Lindberg, David B Rose, Erik P Garrison, Gabor T Marth, Aaron R Quinlan, and Ira M Hall. "SpeedSeq: Ultra-Fast Personal Genome Analysis and Interpretation." Nature Methods 12, no. 10 (October 2015): 966–68. https://doi.org/10.1038/nmeth.3505.

https://github.com/hall-lab/svtyper